Vector-quantized Image Modeling with Improved VQGAN

Year: 2021 | Citations: 660 | Authors: Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang

Abstract

Pretraining language models with next-token prediction on massive text corpora has delivered phenomenal zero-shot, few-shot, transfer learning and multi-tasking capabilities on both generative and discriminative language tasks. Motivated by this success, we explore a Vector-quantized Image Modeling (VIM) approach that involves pretraining a Transformer to predict rasterized image tokens autoregressively. The discrete image tokens are encoded from a learned Vision-Transformer-based VQGAN (ViT-VQGAN). We first propose multiple improvements over vanilla VQGAN from architecture to codebook learning, yielding better efficiency and reconstruction fidelity. The improved ViT-VQGAN further improves vector-quantized image modeling tasks, including unconditional, class-conditioned image generation and unsupervised representation learning. When trained on ImageNet at \(256\times256\) resolution, we achieve Inception Score (IS) of 175.1 and Fr'echet Inception Distance (FID) of 4.17, a dramatic improvement over the vanilla VQGAN, which obtains 70.6 and 17.04 for IS and FID, respectively. Based on ViT-VQGAN and unsupervised pretraining, we further evaluate the pretrained Transformer by averaging intermediate features, similar to Image GPT (iGPT). This ImageNet-pretrained VIM-L significantly beats iGPT-L on linear-probe accuracy from 60.3% to 73.2% for a similar model size. VIM-L also outperforms iGPT-XL which is trained with extra web image data and larger model size.